CSCI 620/Section 01/Mior, Introduction to Big Data, Fall 2201

Assignment 6 – Data cleaning

Description

An additional data file is provided on myCourses which contains, for each line, a JSON document describing movies and other information related to them extracted from Wikidata (https://www.wikidata.org/).

Your tasks

1. Briefly describe the information stored in the file.
   (10 points)

2. Add the information provided by the data file to your MongoDB database of movies. Specifically, you need to include the information about the box office revenue in US dollars, cost of the movie, distributor, and rating. You can choose any value if there are duplicates. Provide a program to perform this operation and report how many successful updates you were able to perform with the given data.

   (Hint: The documents in the file contain the IMDB ids of the movies.)
   (35 points)

3. Assume that the documents in the file did not contain the IMDB IDs, how would you perform the matching process? Provide a program that analyzes the issues you may find in such a scenario. Consider how many values in the new data set match existing documents and whether there are any cases where new values match more than one existing document.
   (25 points)

4. Provide programs that extract the following data to be plotted from MongoDB. Use your favorite visualization program or library to plot and report your results.
   (10 points each)

   4.1. For each genre, a five-number summary of the average ratings of movies with more than 10K votes.
   4.2. Average number of actors per movie by genre as a bar chart for all movies with any actors (i.e. skip documents with no “actors” field).
   4.3. Number of movies produced each year (startYear) as a time series plot.